Interspeech 2024
1-5 September 2024, Kos, Greece

Prompt Tuning for Speech Recognition on Unknown Spoken Name Entities

Xizi Wei, Stephen McGregor

action.ai, London, UK

xizi@action.ai, stephen@action.ai

Abstract

This paper explores the challenge of recognising relevant but previously unheard named entities in spoken input. This scenario pertains to real-world applications where establishing an automatic speech recognition (ASR) model trained on new entity phrases may not be efficient. We propose a technique that involves fine-tuning a Whisper model with a list of entity phrases as prompts. We establish a task-specific dataset where stratification of different entity phrases supports evaluation of three different scenarios in which entities might be encountered. We focus our analysis on a seen-but-unheard scenario, reflecting a situation where only textual representations of novel entity phrases are available for a commercial banking assistant bot. We show that a model tuned to anticipate prompts reflecting novel named entities makes substantial improvements in entity recall over non-tuned baseline models, and meaningful improvements in performance over models fine-tuned without a prompt.

Index Terms: speech recognition, prompt tuning, named entity recognition, out-of-vocabulary words recognition

1. Introduction

Research into computational frameworks for hearing and interpreting names spoken by human voices has tended to focus on the high-level question of whether an automatic speech recognition system can successfully interpret any given voice saying a particular variety of named entities, typically with the assumption that entities are drawn from a set of phrases observed at the time of model training. The research presented in this paper is motivated by a practical extension of this line of work: how do we explore the ability of a model to hear new names said by new speakers, and what modelling configuration offers an optimal solution to this challenging mode of named entity recognition? This question is particularly pertinent to commercial applications of automatic speech recognition, where developers can expect to require quick implementation of support for new users saying likewise new named entity phrases that need to be successfully transcribed for subsequent classification.

There have been many attempts at named entity recognition (NER) from speech via a two-step approach [1] where an STT model converts speech to text and then passes this text along to a semantic NER classifier, or alternatively where an E2E ASR model attempts to directly apply named entity tags to predicted transcriptions [2, 3, 4, 5]. However, these efforts have not considered the specific problem of unknown words in named entities.

As for unknown word recognition, researchers in [6] proposed to use training on synthetic audio for Out-of-Vocubulary (OOV) words to build their E2E ASR system with improved recognition accuracy on OOV words. This approach is compelling, but involves retraining an entire ASR system whenever new named entities are introduced to the vocabulary, which would be time consuming and simply not viable for commercial developers faced with providing almost instant support for new phrases associated with for instance new users or updated product details.

In a context of commercial banking voice assistants, which we investigate here, it is not possible to know the specific names that a speaker is going to use for each turn, but a textual list of potential account names may be available for any given customer. We explore fine-tuning a model to expect prompts that include a list of previously unheard entity phrases. We also explore the sensitivity of the model to long distance prompts by changing the position of the target names in the prompt and also the length of the prompt itself.

In order to carry out experiments on the identification of novel named entities, we have developed a dataset reflecting utterances spoken by customers in a commercial banking context, where we can anticipate that speakers might use the proper names of accounts with which they transact. We consider three different named entity scenarios that might arise in this setting:

- **Seen-and-Heard:** Where we encounter a potentially new voice saying known names (as in names that are likewise present in spoken training data);
- **Seen-but-Unheard:** Where we have access to a list of target named entities, but do not have training data reflecting any voice speaking the name we are trying to hear;
- **Unseen-and-Unheard:** Where we are trying to hear completely unfamiliar names.

In this paper, we explore ways in which an ASR model can be primed in the seen-but-unheard scenario in particular, where in a commercial banking context we might for instance have a list of specific accounts associated with a specific customer. We experiment with a technique for introducing textual prompts to a Whisper [7] model, in the form of lists of values of entities that could possibly occur in a spoken input. We explore the impact of training a model to explicitly anticipate prompts corresponding to possible entity values. We show that prompt-tuning the model with a list of anticipated entity phrases improves the recall rate for recognising the seen-but-unheard named entities by 0.433 compared to only giving a prompt during decoding with the baseline Whisper model, by 0.015 over performance of a model fine-tuned but not trained to anticipate prompts, and by 0.284 over a fine-tuned model that is not prompted at all. Furthermore, WERs on full transcription is also consistently decreased using our approach.

In what follows, Section 2 discusses related work on prompting in both natural language processing and speech pro-

10.21437/Interspeech.2024-1162

cessing. Section 3 introduces the dataset we have developed for this research. Section 4 and Section 5 discuss the methodology and experiments with results, followed by conclusions.

2. Related Work

In the natural language processing (NLP) field, large language models trained on a sufficiently large amount of data have been shown to perform well across many domains and tasks [8]. In this area, "prompting" in its broadest sense refers to providing a language model with the beginning of a sequence of tokens to be predicted, with the expectation that the most likely continuation of this sequence can be treated as output for a particular task. Through the definition of a new prompting function a pre-trained language model is able to perform few-shot or even zero-shot learning, adapting to new scenarios with little or no labelled data [9]. Prompt learning has proved to be particularly effective when adapting language models to novel domains [10]. Researchers have also demonstrated that fine-tuning language models with instructions improves zero-shot performance on unseen tasks [11].

In the field of speech processing, the encoder-decoder transformer [12] based model Whisper has been trained on a large amount of labelled multilingual and multitask audio data, and has been shown to generalise well across domains, tasks, and languages [7]. Whisper is designed with a position allowing for the introduction of prior context for conversational or long-range transcription tasks, where this context is the text of speech immediately preceding an audio input, so playing a similar role to the application of prompting in transformer based language modelling. This reference text is concatenated to the beginning of decoder input: during fine-tuning of a model, the decoder mostly captures domain-specific information from the reference text while the encoder learns audio-related features.

Prompt-tuning with Whisper has been shown to improve performance on target-speaker ASR [13], and out-of-domain ASR for spoken language assessment [14], where trainable soft prompts were utilised. Researchers in [15] furthermore showed that tuning the Whisper model with a text prompt that contains textual domain information can improve the recognition accuracy for target domain, which suggested that a prompt-tuned whisper learns to utilise the information in the given prompt. These techniques have, however, leveraged context which is either abstract (in the case of soft prompt vectors) or general (in the case of a domain description).

3. Data

We have endeavoured to conceive a scenario in which a developer may have an ASR model that is expected, on short notice, to be capable of classifying and transcribing previously unanticipated names. With this goal in mind, we have generated 7,031 unique utterances pertaining to commercial banking. The focus of these utterances is enquiries about balances and transactions, including questions about incoming and outgoing payments to a variety of named payers and payees, broken down as follows:

- 2401 utterances including at least one account name, for instance "search the overdue invoices for **Alpha Limited**";
- 4630 utterances do not involve an account name , for instance "I would like to know my balance".

This data features 378 unique account names occurring across the 2401 account name utterances. We segregated all utterances into five clusters designed to have roughly equal distribution of account names. Any given account name will appear in only one cluster, enabling cross-evaluation of models trained on four clusters, including account names in those clusters, and then evaluated on names only present in the fifth held-out cluster.

We recruited 63 speakers based in the UK to read our scripted utterances over a telephone line. The least frequent speaker read 5 utterances, while the most frequent read 180; most speakers (49) read between 80 and 160 utterances. Each of the 63 speakers read utterances from only one of the five account name clusters, meaning that a novel account name encountered in a held-out evaluation cluster will be spoken by a voice that is likewise not present in training data. We have had each utterance in our dataset transcribed, taking into account any deviations from the script used to generate the utterance. In total, the dataset has 12.7 hours of 8kHZ transcribed telephone audio data, upsampled to 16kHZ for experiments with Whisper.

In order to test different modelling configurations, we have stratified our training and testing data in two different ways. In the first arrangement, we maintain the separation between different account names: this set-up is used to explore both the unseen-and-unheard and seen-but-unheard evaluation objectives. In the second arrangement, we redistribute the stratification of different speakers so that the same account names occur in different evaluative folds, creating a basis for exploring the seen-and-heard scenario. In each configuration, the count of overall utterances as well as the count of those featuring named entities are distributed roughly equally across folds, while any given speaker only appears in one fold.

We have also generated an evaluation dataset in which our human testers call a voice banking bot playing the role of customers enquiring about their bank accounts and interacting with the system in an unscripted manner. We denote this as "unscripted data". The unscripted data offers a more challenging transcription objective than the cross-evaluation data, as it reflects spontaneous speech with natural features like disfluency, hesitation, and repair, and may contain background noise. This dataset contains 2,520 utterances in total, with 655 utterances including at least one account name.

4. Methodology

Here we describe a methodology for applying a prompt containing an explicit rendition of a potential portion of a target transcription, in particular a seen-but-unheard named entity phrase.

Whisper's design allows for a `<|startofprev|>` token used to indicate the beginning of previous context, with a `<|startoftranscript|>` then used to indicate the end of context and the beginning of a predicted transcription. We place a list of potential seen-but-unheard named entity values in the position between these tokens, concatenated by ', '.

We have run experiments exploring the application of entity prompts to both base and domain-specific fine-tuned Whisper-small models. In the case of fine-tuned models, we used a batch size of 8 and a learning rate of 1e-5 for all training, and saved the model at 1000 steps. While computing loss during prompt-tuning, the prompt tokens were ignored; the model was trained to predict only the transcription tokens.

As described in section 3, there are 4,630 utterances that do not include an account name and 2,401 utterances including at least one account name. In order to explore the best strategy for fine-tuning the model to only predict named entities from the given name list prompt when the context is likely to have a name, we have developed three different prompt formats for prompt-tuning:

- **PT_1** Supply every utterance with an entity prompt list during training, regardless of whether or not the target transcription involves a named entity. This is closest to the real world scenario (and also our evaluation scenario) where we won't know whether an utterance will contain an account name or not.
- **PT_2** Only use a prompt during training when there is a named entity in the target transcription, and otherwise provide an empty list as a prompt, in order to explore if the model can learn to ignore the prompt altogether when the overall context of the target transcription makes the presence of a named entity unlikely.
- **PT_3** Provide a prompt that contains the target entity phrase for a specific transcription in random position, and leave the rest of the list empty when there is an entity in the target transcription; for non-entity target transcription, leave the prompt list empty. Training with this strategy might help the Whisper decoder learn to pay more attention to target names in a list of prompted entities at inference time.

During all modes of training, the order of entities in a prompt list is randomly shuffled. For evaluation, each input utterance includes a full list of named entities as a prompt. We have also experimented with putting target names at the beginning of the named entity lists versus shuffling the names in the name lists. The shuffled prompt condition is more in line with what will be encountered in real world scenarios.

5. Experiments

We have run a series of experiments to investigate the impact of named entity prompting, with and without prompting included as a feature of model tuning.

5.1. Evaluation metric

The focus of our experiments is the accuracy of the transcription of any named entities in an utterance, and also the overall accuracy of the transcription. As such we measure model performance by the recall rate of correctly recognised named entities and Word Error Rate (WER) of full transcriptions. Given the actual target names used for each utterance as reference, we compute recall by counting the times a correct entity name occurs in the ASR prediction for each utterance and dividing this count by the total number of expected occurrences of entity names in the reference. If an entity name has been utilised more than once in a particular utterance, it will be counted as one occurrence.

5.2. Results

Table 1 presents results from a cross-evaluation of the Whisper model in different entity scenarios without any prompting. The Whisper baseline shows a low recall rate on recognising named entities and a high WER. This is a consequence of the lack of domain-specific tuning and corresponding exposure to the vocabulary and syntax associated with our commercial banking target data, as well as differences in the acoustic conditions of our telephone data as compared to Whisper's base training data.

Fine-tuning (FT) the model with in-domain data in seen-and-heard and unseen-and-unheard scenarios significantly improves performance. However these results also clearly indicate that it is more challenging for the model to recognise named entities that have never occurred in the training data, as indicated by recall in particular in the unseen-and-unheard scenario.

Table 1: *Cross-evaluation of the fine-tuned models for Spoken Named Entity Recognition without prompting.*

Model	Recall	WER%
Whisper baseline	0.078	32.01
FT Seen-and-Heard	0.802	6.38
FT Unseen-and-Unheard	0.550	7.44

As mentioned in Section 3, stratification for the seen-and-heard scenario involves distributing the same named entities across all evaluative folds, while stratification for the unseen-and-unheard scenario (as well as the seen-and-unheard scenario) involves keeping all instances of a given named entity in the same fold for evaluation.

Table 2 shows results for cross-evaluations in which lists of five names are used as prompts, comparing applications of prompts to baseline models, models fine-tuned without prompting, and models fine-tuned with prompting. The data stratification strategy for cross-evaluation is the same as with the unseen-and-unheard scenario, but a textual five-name list that includes any target names is given as prompt, and so these results reflect the seen-but-unheard scenario, as the target entities are "seen" in the form of a prompt. Results for the three different strategies for presenting prompts during prompt-tuning are shown here, corresponding to the descriptions of PT_1, PT_2, and PT_3 in Section 4. We also compare a prompting strategy where an in-transcription target entity is always provided at the beginning of a list of prompts, indicated in Table 2 as "Begin", to a strategy where a target entity is placed at random in the list, indicated as "Shuffle".

Table 2: *Cross-evaluation of **Seen-but-Unheard** Spoken Named Entity Recognition with five-name lists as prompts for decoding on baseline Whisper, fine-tuned Whisper, or prompt-tuned Whisper (PT_1, PT_2, and PT_3). The target account names are at the beginning of the name list or in random positions.*

Model	Begin		Shuffle	
	Recall	WER%	Recall	WER%
Whisper baseline	0.412	24.89	0.401	26.41
FT Whisper	0.763	6.57	0.819	6.72
PT_1	0.776	6.20	**0.832**	**5.96**
PT_2	0.775	5.97	**0.832**	**5.89**
PT_3	0.780	6.02	**0.834**	**5.76**

Our results show that prompt-tuning the model with a list that includes anticipated named entities leads to better performance than merely fine-tuning the model with in-domain data without any prompts. With the best performing prompting strategy PT_3, the prompt-tuned models achieved an improvement on recall rate on recognising named entities from the given shuffled name list of 0.015 and a decrease of 0.96% on full transcription WER as compared to a model that was simply fine-tuned and not prompt-tuned. Furthermore, providing a prompt to the baseline Whisper model during decoding increases the recall rate from 0.078 to 0.412 and reduces the WER from 32.01% to 24.89%. This suggest that even the Whisper baseline model has some sensitivity to a given prompt, but prompting is still significantly improved through fine-tuning.

Table 3 shows the ASR outputs for one utterance from a fine-tuned model (denoted as "FT"), providing a prompt to the fine-tuned model (denoted as "FT + prompt"), and providing a prompt to the prompt-tuned model (denoted as "PT_3 + prompt"). The first row is the named entities list given as a

Table 3: *Error analysis on the ASR hypothesis for one utterance from fine-tuned model, giving a prompt to the fine-tuned model, and giving a prompt to the prompt-tuned model.*

Prompt	***fortuna and clark***, l b a, zacaris, ***manzell consultant***, l b f
Reference	did i definitely pay ***fortuna and clark*** last saturday can you see if any money has come in from the ***manzell consultant*** also find out how my balance is doing
FT	did i definitely pay *for tuna and club* last saturday can you see if any money has come in from the *mandible consultants* also find out how my balance is doing
FT + prompt	did i definitely pay *for tuna and* ***clark*** last saturday can you see if any money has come in from the *mansill consultant* also find out how my balance is doing
PT_3 + prompt	did i definitely pay ***fortuna and clark*** last saturday can you see if any money has come in from the ***manzell consultant*** also find out how my balance is doing

prompt to the models for this particular utterance. The target names are marked in bold. From the output, we can see that the FT model achieves good accuracy on generic words, but failed to recognise the unseen-and-unheard named entities here. An FT model prompted during decoding recovers from some errors involving substitution of an entity name with a generic word, but still fails in particular on getting precise spellings of entiy names. With the "PT_3 + prompt" method, the model successfully utilises the information from the given prompt and outputs the target named entities precisely.

It is more challenging for both fine-tuned and prompt-tuned models to correctly recognise the target name if the target name is at the beginning of the given name list, as indicated by results in the "begin" condition in Table 2. One interpretation of this is that the attention mechanism within the decoder may have less sensitivity over longer contexts, as the beginning of the prompt list will be further from the target transcription in the modelled sequence of tokens.

To further investigate the model's sensitivity as a function of prompt length (and so potentially underlying aspects of attention), we evaluate the fine-tuned and prompt-tuned model on the 2,000 utterances that only contain one account name by placing the target entity name in 4 different positions in a 15-name entity prompt list. For this experiment, the models are trained with 15-name shuffled lists. Figure 1 shows how recall and WER changes when target entities are placed in positions 1, 5, 10, and 15 in the 15-name prompt list, where position 1 is the beginning of the list and so the farthest from the beginning of the transcription. We can see that performance tends to improve with target names closer to the end of the list and the beginning of the transcription. This suggests that the model's attention starts to decay over distance from the transcription itself, particular once it is further than 10 account names away.

Table 4 shows results for models trained on scripted data and evaluated against unscripted data, so data reflective of more natural conversational inputs, as described in Section 3. For prompt-tuning, we use the optimal strategy from experiments on scripted data, so PT_3 with shuffled target entities. During inference, we use five-name lists that include the target name in random positions as prompts for the baseline, fine-tuned, and prompt-tuned models. Here we can again see the benefit from priming the model for seen-but-unheard named entity recognition with prompt-tuning, with an improvement of 0.028 in recall rate compared with no-prompt fine-tuned models.

6. Conclusion

We have posed a question about how to correctly hear named entities in a situation where we can anticipate coming across

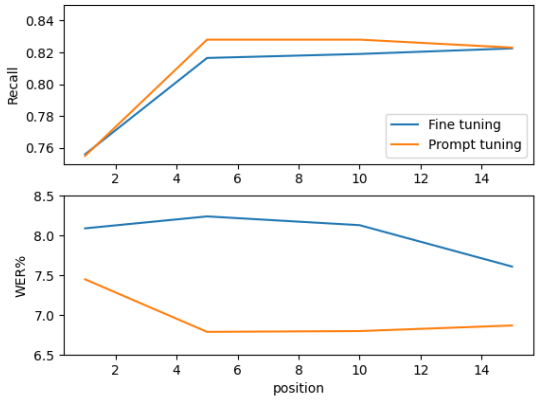

Figure 1: *Fine-tuned and Prompt-tuned models performance when given a 15-name list prompt with target name in different positions.*

Table 4: *Evaluation on unscripted speech with 5 shuffled name list as prompt. Model trained on scripted data with the optimal prompting stragety.*

Model	Recall	WER%
Whisper baseline	0.637	46.09
FT Whisper	0.869	7.84
PT_3	0.897	7.61

previously unencountered entity phrases. In order to explore this question, we have developed a dataset that isolates named entities in different folds for the purpose of comparatively cross evaluating what we've labelled seen-but-unheard and unseen-and-unheard scenarios. We've proposed a novel technique for a model which supports a seen-but-unheard scenario, involving prompting a Whisper model with anticipated and so "seen" entity phrases not "heard" during in-domain tuning of a model.

Our results show the effectiveness of our strategy, with models that are trained to anticipate entity prompts notably outperforming models that are prompted but not trained to expect prompts, and even more significantly outperforming models that are not prompted at all. This technique for priming a model to be prompted offers a promising solution to a real-world scenario where expected named entity phrases become available at close to the point of transcription prediction, as a prompt list can be instantaneously provided to an appropriately tuned model.

Our analysis of model performance suggests that there is a relationship between the position of target names in a prompt list and the ability of a model to identify those names in a target transcription. This observation invites further speculation about the attention mechanism of the language modelling performed by the decoder component of a Whisper model. To some extent the model must be learning to associate a higher probability with a repetition of a token seen near the beginning of a sequence (in the prompt provided in the prior context portion of decoder input) and further on in the sequence (in the portion of decoder output corresponding to a predicted transcription). Further experimentation designed to explore the nature of the relationship between context and prediction as well as the underlying model features could point the way towards further advances in our prompt-tuning strategy.

7. References

[1] I. Cohn, I. Laish, G. Beryozkin, G. Li, I. Shafran, I. Szpektor, T. Hartman, A. Hassidim, and Y. Matias, "Audio de-identification - a new entity recognition task," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*. Association for Computational Linguistics, pp. 197–204.

[2] S. Ghannay, A. Caubrière, Y. Estève, A. Laurent, and E. Morin, "End-to-end named entity extraction from speech," in *CoRR*, 2018.

[3] H. Yadav, S. Ghosh, Y. Yu, and R. R. Shah, "End-to-end named entity recognition from english speech," *Proc. INTERSPEECH*, 2020.

[4] A. Pasad, F. Wu, S. Shon, K. Livescu, and K. J. Han, "On the use of external data for spoken named entity recognition," 2022.

[5] S. Shon, A. Pasad, F. Wu, P. Brusco, Y. Artzi, K. Livescu, and K. J. Han, "Slue: New benchmark tasks for spoken language understanding evaluation on natural speech," in *ICASSP*, 2022.

[6] X. Zheng, Y. Liu, D. Gunceler, and D. Willett, "Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end asr systems," in *ICASSP*, 2021.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:160025533

[9] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, jan 2023.

[10] E. Ben-David, N. Oved, and R. Reichart, "Pada: Example-based prompt learning for on-the-fly adaptation to unseen domains," in *TACL*, 2022.

[11] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=gEZrGCozdqR

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[13] H. Ma, Z. Peng, M. Shao, J. Li, and J. Liu, "Extending whisper with prompt tuning to target-speaker asr," in *ICASSP*, 2024.

[14] R. Ma, M. Qian, M. Gales, and K. M. Knill, "Adapting an asr foundation model for spoken language assessment," in *9th Workshop on Speech and Language Technology in Education (SLaTE)*. ISCA, 2023.

[15] F.-T. Liao, Y.-C. Chan, Y.-C. Chen, C.-J. Hsu, and D. shan Shiu, "Zero-shot domain-sensitive speech recognition with prompt-conditioning fine-tuning," in *ASRU*, 2023.